Classification of Health Webpages as Expert and Non Expert with a Reduced Set of Cross-language Features

Natalia Grabar[1,2], PhD, Sonia Krivine[3], MS, Marie-Christine Jaulent[1], PhD

[1]INSERM, UMR_S 872, Eq. 20, Paris, F-75006 France; Université René Descartes, Paris, F-75006 France
[2]Health on the Net Foundation, SIM/HUG, 24 rue Micheli-du-Chrest, Geneva, Switzerland
[3]FircoSoft, 37 rue de Lyon, 75012 Paris, France

Abstract

Making the distinction between expert and non expert health documents can help users to select the information which is more suitable for them, according to whether they are familiar or not with medical terminology. This issue is particularly important for the information retrieval area. In our work we address this purpose through stylistic corpus analysis and the application of machine learning algorithms. Our hypothesis is that this distinction can be performed on the basis of a small number of features and that such features can be language and domain independent. The used features were acquired in source corpus (Russian language, diabetes topic) and then tested on target (French language, pneumology topic) and source corpora. These cross-language features show 90% precision and 93% recall with non expert documents in source language; and 85% precision and 74% recall with expert documents in target language.

Introduction

When searching the Web, eight out of ten users look for online health information.[1] The information found presents different technical levels: documents can be more or less difficult to understand to non expert medical users depending on topics presented and terms and words used. The existence of this technical heterogeneity is not transparent. Yet it should be clearly indicated, especially for the non expert users, as this situation can have direct impact on users' healthcare or communication with medical professionals.[2,3] For this reason, search engines should propose solutions for the distinction of document types according to whether they are written for medical experts or non expert users. Notice that medical portals like HON (*www.hon.ch*), CISMeF (*www.chu-rouen.fr/cismef*), or general search engine Google Coop for health (*www.google.com/coop*), propose this distinction but it is based on the manual categorisation of webpages and websites.

Among existing automatic approaches, let's quote: (1) linguistically founded formulae (*i.e.*, Flesch,[4] Fog,[5] Lix[6]), which rely on criteria like average length of words and sentences; (2) combination of these formulae with specialised medical terminologies[7] in order to take into account the medical dimension; (3) application of the text categorisation algorithms to various features: manually weighted MeSH terms,[8] *n-grams* of characters,[9] combination of linguistic features, word difficulty and unigrams,[10] documents' vocabulary.[11] These experiments show interesting results but the length of linguistic units is not systematically correlated with their difficulty, and deciphering features or building learning corpora can become a tedious task.

We work in a multilingual context and aim at distinguishing expert and non expert medical documents in different laguages (French, Russian, Japanese, English). To ease this task, we propose to use a small set of features, which would be easy to define and to apply to any new language or domain. Assuming that content and style of documents represent the context of their creation and usage[12] (*i.e.*, addressee, aim when creating a document), we propose to set features at the stylistic level. They are defined on the basis of source corpus and then applied to the target corpus. Languages and domains of these two corpora are different. The purpose of this work consists in selecting a small set of features and applying them through machine learning algorithms.

Material

We distinguish source and target corpora. Source corpus contains documents on diabetes in Russian, target corpus contains documents on pneumology in French. The source corpus has been built through general Russian-speaking search engines (*www.google.ru*, *www.yandex.ru*, *www.rambler.ru*, *www.aport.ru*) which were queried with keywords related to *диабет и питание* (*diabetes and diet*). The distinction between expert and non-expert documents has been made manually, by a non expert medical evaluator. Webpages for French corpora have been detected through the specialised search engine of the CISMeF portal. Referenced documents are indexed by librarians with MeSH which allows to reach documents related to a specific medical area. We used keyword *pneumologie* (*pneumology*) when querying CISMeF. For the distinction between expert and non expert

	Russian			French		
	S	D	O	S	D	O
Expert	21	35	116,000	46	186	371,045
Nexpert	52	133	190,000	31	80	87,177
Total	69	168	306,000	58	266	458,222

Table 1: Expert and non expert corpora in Russian and French languages.

documents, we used *type de ressource* (*resource type*) annotations proposed by this portal. This annotation can have various values (*i.e.*, *course material*, *guidelines*, *information for patients*). We grouped relevant ones into two aimed categories: expert and non expert.

On the basis of the collected URLs we downloaded their content with the `wget` tool. Documents are originally encoded with different character sets (*i.e.*, win1251, iso-8859-5, koi-8r, iso-latin1). When possible, they were converted to a common encoding utf8. Documents are available in text and HTML formats.

Table 1 indicates size and composition of studied corpora: *S* stands for number of sources, *D* for number of documents and *O* for number of occurrences in each corpus. French corpus contains more documents but they have been collected on fewer number of sites. This is certainly due to the current Internet situation for these languages: in French some sites are specialised in providing health information, while in Russian such documents are spread over the web. Moreover, we can observe a difference between sizes of expert and non expert corpora: non expert corpus is bigger in Russian, while expert corpus is bigger in French. These corpora are used by machine learning algorithms during learning and test steps.

Methods

We use several machine learning algorithms (`Naive Bayes`, `J48`, `RandomForest`, `OneR` and `KStar`) in order to compare their performance and to test the consistency of the established set of features. The main challenge of the method relies on the universality of proposed features defined on the basis of the source corpus and then applied to the target corpus, both composed of documents related to different domains and languages.

Feature selection

Stylistic features have emerged from a previous contrastive study of expert and non expert corpora in Russian[13]. Use of lexicometric and NLP tools (Lexico3 and Unitex) permitted to discriminate them. For the current work, we selected a set of 14 features related to the document structure, personal pronouns, punctuation marks and uncertainty.

Document structure and layout. Among HTML tags used for the structuring and layout of documents we discriminated tags for displaying: images <img>, tables <table>, lists <ul> and <ol>, hypertext links <a>, italic <i> and bold <b> text. Among these structural features, <img>, <table> and <a> appear to be specific to non expert, and <ul>, <ol>, <i> and <b> to expert documents.

Personal pronouns. The general assumption with personal pronouns is that they are specific to non expert documents. Indeed, in these documents, authors are expected to address directly their addressees, while scientific documents should remain impersonal. We studied four pronouns: 1^{st} and 2^{nd} singulars, and 1^{st} and 2^{nd} plurals. They are all specific to non expert documents. When some of these pronouns occur in scientific documents their use is related to direct citations, *i.e.* questions which would be asked to patients.

Punctuation marks. Punctuation marks can reflect the complexity of sentences (comas, dots, colon, semicolon, parentheses, etc.), give indication on emotions (question and exclamation marks), introduce citations (quotation marks), etc. We take into account question and exclamation marks, which are specific to non expert documents.

Uncertainty modality. Russian uncertainty modality *бы* (*/by/*) and French conditional modality of verbs (both being close to English auxiliar verbs *should*, *would* and *could*) are supposed to be specific to non scientific documents.

Evaluation

Learning and test are performed on independent corpora, composed of respectively 66% and 33% of the whole corpus collected. Evaluation is done through the precision, recall, F-measure and error rate.

Results and Discussion

We used the set of 14 stylistic and structural features and Weka[14] (*Waikato Environment for knowledge analysis*) tool for providing the machine learning algorithms. We used its default parameters for five algorithms (`NaiveBayes`, `J48`, `RandomForest`, `OneR` and `KStar`) representing different families of classifiers.

Quantitative evaluation

Corpus in Russian, composed of 168 documents, has been split into two independent subcorpora: learning corpus (66%: 110 documents) and test corpus (33%:

Method	Expert			Non expert			Err
	P	R	F	P	R	F	
NBayes	43	83	57	94	72	82	26
J48	**83**	**42**	56	86	98	92	14
RForest	**83**	**42**	56	86	98	92	14
OneR	43	25	32	82	91	87	22
KStar	70	58	64	**90**	**93**	91	14

Table 2: Evaluation of algorithms on Russian corpus

Method	Expert			Non expert			Err
	P	R	F	P	R	F	
NBayes	93	36	52	31	91	46	50
J48	81	83	82	43	41	42	27
RForest	**87**	**81**	84	**52**	**64**	57	23
OneR	83	87	85	53	45	49	23
KStar	85	74	79	42	59	49	30

Table 3: Evaluation of algorithms on French corpus

58 documents). Results obtained are presented in table 2. For each method (first column), we indicate figures related to its performance: precision, recall, F-measure and error rate. `KStar` shows the best results with *non expert* documents: 90% precision and 93% recall, and nearly the best results for the *scientific* category: 70% precision and 58% recall. `J48` and `RandomForest`, both using decision trees, present identical results for two studied categories: 83% precision and 42% recall with *scientific* documents and 86% precision and 98% recall with *non expert* documents. From the point of view of precision, these two algorithms are suitable for the categorisation of documents as *scientific*. Such results can be considered as satisfying even if they show a low recall with *scientific* documents. This weakness, observable with all the methods, can be explained by the small learning set of expert documents in Russian. Two remaining algorithms, `NaiveBayes` and `OneR`, have generated error rate of over 20% and a low precision (43%) within the *scientific* category.

Table 3 indicates evaluation results of the same algorithms applied to French corpus (175 documents for learning and 91 for test). `RandomForest` has generated the most competitive results for both categories (*expert* and *non expert*). Surprisingly, `OneR`, based on the selection of only one rule, produced results which are close to those of `RandomForest`. Error rate is important with `NaiveBayes` and `KStar`. Here again, the size of corpora seems to impact the results: scientific corpus, which is larger than non expert corpus, provides better categorisation.

Among the most efficient algorithms, we notice `J48`, `RForest` and `KStar` for Russian; and `RForest` for French. `NaiveBayes` shows low performance in both corpora.

Qualitative evaluation

We present and discuss the following issues: generated language models, errors common to different algorithms, analysis of an ambiguous Russian document and suitability of proposed cross-language features.

Language model. Language models generated by `OneR` and `J48` have been analysed. `OneR` selects one (best) rule in each corpus. In our experiment, this algorithm selected hypertext link <a> tag in Russian and 2^{nd} plural pronoun in French. These features allow to produce nearly the best results in the target corpus (French), while in Russian this algorithm is the least competitive.

The model produced by `J48` in Russian selects hypertext link <a> tag together with 1^{st} singular pronoun *я* (*I*), italic characters (tag <i>), lists (<ol>) and table (<table>) tags. On French corpora, `J48` selects the following five features: 2^{nd} plural pronoun, <table> tag, 2^{nd} singular pronoun, <ol> tag and exclamation mark. `J48` is one of the most suitable algorithms in Russian but it shows moderate performance in French. Only two of the selected features are common to the both studied corpora: <ol> and <table> tags.

As noticed, relevance of this set of features has been first tested with lexicometric tools on Russian corpus. Current work allows to better weight and compare their performance in both source and target corpora. Like in a previous study,[15] the present results indicate that even a reduced set of features can be adequate for text categorisation, for instance for making distinction between expert and non expert documents. Surprisingly, several relevant features are related to the `HTML` tagging of documents, which suggests that categorisation of web documents should be based on textual as well as on non textual criteria. According to the theory of genres,[16] this observation emphasizes the importance of documents' layout, typography and intertextuality when analysing them according to their genres and discourses.

Analysis of errors common to various classifiers. `J48` algorithm generated 17 wrong categorisations in Russian and, among them, six are also wrongly categorised by other applied algorithms. Among these 6 documents:

- 4 have been manually labelled as *non expert* while the automatic system assigns them to the *expert* category;
- 2 documents have been considered as ambiguous

when manually categorising them and even excluded from the Russian corpus during the previous work on definition of the set of features.[13] In the current work, these documents were part of the test corpus.

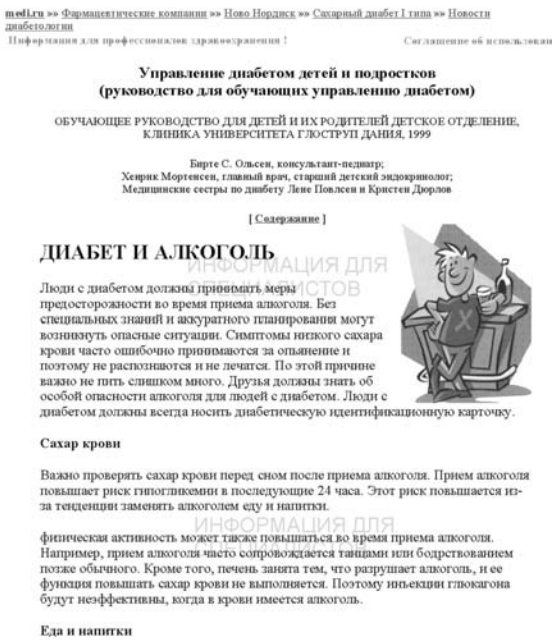
This observation suggests that some documents may be ambiguous when categorising them through a manual or automatic process. Moreover, this observation indicates that discourse distinction between expert and non expert document is set on a continuum axis, and that there is no dichotomy between them.

Analysis of an ambiguous document. As noticed, during the manual categorisation of Russian documents, some of them presented some difficulty and have been excluded from corpus because of the diversity of expert and non expert features they contained. Figure 1 presents such a document:

- This document contains the following *expert* features: (1) its layout, (2) presence of title, (3) its location in a medical portal within the directory *Information for health professionals* together with scientific papers, (4) mention of authors and of their institutions, (5) presence of the watermark *Информация для специалистов* (*Information for experts*).
- This document presents the following *non expert* features: (1) presence of colour image, (2) use of 2^{nd} singular pronoun (3) and of imperative forms of verbs. The use of pronouns is due to the fact that this document is a guideline entitled *Диабет и алкоголь у подростков* (*Diabetes and alcohol and teenager*) written for teenagers with diabetes, where advises are written with 2^{nd} singular pronoun, certainly so that young people feel more concerned with this guide.

We applied the categorisation system to this document. Three classifiers (`NaiveBayes`, `RandomForest` and `OneR`) categorised it as *non expert*, and two classifiers (`J48` and `KStar`) as *scientific*. Such results are interesting as they again highlight the real difficulty to assign some documents to discourse related categories. As noticed, this difficulty can appear with both manual and automatic approaches. Thus, when several methods are applied, their voting can be used to help the decision making about the document category.

Suitability of the proposed features. The proposed reduced set of features contains 14 criteria related to the document structure, personal pronouns, punctuation and uncertainty marks. Obtained results seem to indicate that these stylistic features are suitable for the categorisation of documents according to their discourse



Управление диабетом детей и подростков
(руководство для обучающих управлению диабетом)

ОБУЧАЮЩЕЕ РУКОВОДСТВО ДЛЯ ДЕТЕЙ И ИХ РОДИТЕЛЕЙ ДЕТСКОЕ ОТДЕЛЕНИЕ, КЛИНИКА УНИВЕРСИТЕТА ГЛОСТРУП ДАНИЯ, 1999

Бирте С. Ольсен, консультант-педиатр;
Хенрик Мортенсен, главный врач, старший детский эндокринолог;
Медицинские сестры по диабету Лене Повлсен и Кристен Дюрлов

[Содержание]

ДИАБЕТ И АЛКОГОЛЬ

Люди с диабетом должны принимать меры предосторожности во время приема алкоголя. Без специальных знаний и аккуратного планирования могут возникнуть опасные ситуации. Симптомы низкого сахара крови часто ошибочно принимаются за опьянение и поэтому не распознаются и не лечатся. По этой причине важно не пить слишком много. Друзья должны знать об особой опасности алкоголя для людей с диабетом. Люди с диабетом должны всегда носить диабетическую идентификационную карточку.

Сахар крови

Важно проверять сахар крови перед сном после приема алкоголя. Прием алкоголя повышает риск гипогликемии в последующие 24 часа. Этот риск повышается из-за тенденции заменять алкоголем еду и напитки.

физическая активность может также повышаться во время приема алкоголя. Например, прием алкоголя часто сопровождается танцами или бодрствованием позже обычного. Кроме того, печень занята тем, что разрушает алкоголь, и ее функция повышать сахар крови не выполняется. Поэтому инъекции глюкагона будут неэффективны, когда в крови имеется алкоголь.

Еда и напитки

Figure 1: Example of an ambiguous document.

(*expert* and *non expert*). Indeed, their application to target corpus shown promising performance, although the source and target corpora are composed of documents in different languages and from different medical domains. On the basis of this experience, we assume that the proposed features may be indeed used through various languages and domains. Moreover, they are easy to adapt to a new language. But their application in new corpora has to be verified. One of the limitations of these features is that some of them remain specific to HTML documents.

Conclusion and Perspectives

We have presented an experiment on automatic distinction of expert and non expert webpages. Learning algorithms and a set of 14 stylistic features have been used. Features have been acquired on source corpus (Russian language, diabetes related topic) and then applied to target (French language, pneumonology related topic) and source corpora. The cross-domain and especially cross-language aspect of features seems to be a new issue in the text categorisation area.

Evaluation results show that, in our experiment, decision tree algorithms `J48` and `RandomForest` are the most suitable for the categorisation of documents as expert and non expert. They generate the best results in target corpus (up to 87% precision and 81% recall), and for the *expert* category in source corpus (83% precision and 42% recall). Their performance with *scientific* documents from source corpus is less

competitive, which is certainly due to the small size of this corpus. But `KStar` algorithm provides 90% precision and 93% recall for this category in Russian. As noticed, results depend on the size of learning corpora: in Russian, the *non expert* corpus is bigger and produced results are better for this category. In French, the situation is inversed: *expert* corpus is bigger and results produced for this category are better. This can be explained by the fact that larger corpora contain more various, and possibly exhaustive, data: the acquired learning model will be more complete. For this reason, it would be interesting to apply the system to a larger collection of documents and to confirm the efficiency of the acquired language models. But we can consider these results as promising, especially as documents are extracted from various websites and learning and test are performed on independent data.

The results obtained seem as well to indicate that the proposed stylistic features are suitable for the categorisation of documents according to their discourse: their application to target corpus shown promising performance, although the source and target corpora are composed of documents in different languages and describe different medical topics. Nevertheless, it could be interesting to apply other linguistically motivated criteria, for instance detection of argumentation[17] or of simple stopwords.[18] The WEKA's automated feature selection can also be used for this purpose.

We observed the existence of ambiguous documents which are difficult to assign to any of the two categories, and noticed that expert and non expert categories are set on a continuum axis and should not be considered as opposite categories. Use of various classifiers and their voting can be an interesting approach for the detection and possible categorisation of such ambiguous documents.

We assume that an enhancement to the used algorithms can be achieved in the future through their tuning or through the feature selection. Moreover, the obtained results can be improved by distinguishing more specific categories of documents (*i.e.*, cook recipes, articles, food recommandations within *non expert* category). Given the small size of corpora, the n-fold cross-validation would solidify the choice of algorithms.

Additionally, we built an intermediate French corpus composed of documents written for medical students (courses, teaching material). It could be interesting to categorise this material through the proposed language model. It could be also interesting to apply our method to other medical areas and genres, and to compare it with results produced by other approaches.

REFERENCES

1. Fox S. Online Health Search 2006. Most Internet users start at a search engine when looking for health information online. Very few check the source and date of the information they find. Technical report, Pew Internet & American Life Project, Washington DC, 2006.
2. AMA . Health literacy: report of the council on scientific affairs. Ad hoc committee on health literacy for the council on scientific affairs, American Medical Association. *JAMA* 1999;281(6):552–7.
3. McCray A. Promoting health literacy. *Journal of American Medical Informatics Association* 2005;12:152–63.
4. Flesch R. A new readability yardstick. *Journal of Applied Psychology* 1948;23:221–33.
5. Gunning R. *The art of clear writing*. McGraw Hill, New York, NY, 1973.
6. Björnsson H and Härd af Segerstad B. Lix pä franska och tio andra spräk. *Stockholm: Pedagogiskt centrum, Stockholms skolförvaltning* 1979.
7. Kokkinakis D and Gronostaj MT. Comparing lay and professional language in cardiovascular disorders corpora. In: Pham T., James Cook University A, ed, WSEAS Transactions on BIOLOGY and BIOMEDICINE, 2006:429–37.
8. Zheng W, Milios E, and Watters C. Filtering for medical news items using a machine learning approach. In: AMIA, 2002:949–53.
9. Poprat M, Markó K, and Hahn U. A language classifier that automatically divides medical documents for experts and health care consumers. In: MIE 2006, Maastricht. 2006:503–8.
10. Wang Y. Automatic recognition of text difficulty from consumers health information. In: IEEE , ed, Computer-Based Medical Systems, 2006:131–6.
11. Miller T, Leroy G, Chatterjee S, Fan J, and Thoms B. A classifier to evaluate language specificity of medical documents. In: HICSS, 2007:134–40.
12. Zweigenbaum P, Jacquemart P, Grabar N, and Habert B. Building a text corpus for representing the variety of medical language. In: MEDINFO, 2001:290–4.
13. Krivine S, Tomimitsu M, Grabar N, and Slodzian M. Relever des critères pour la distinction automatique entre les documents médicaux scientifiques et vulgarisés en russe et en japonais. In: Mertens P, Fairon C, and Dister A, eds, TALN, Louvain, Belgique. 2006:522–31.
14. Witten I and Frank E. *Data mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco, 2005.
15. Goldstein J and Sabin RE. Using speech acts to categorize email and identify email genres. In: 39th Hawaii International Conference on System Sciences, 2006.
16. Genette G. *Théorie des genres*. Seuil, Paris, 1986.
17. Ruch P, Boyer C, Chichester C, et al. Using argumentation to extract key sentences from biomedical abstracts. *Int J Med Inform* 2006;76(2-3):195–200.
18. Rillof E. Little words can make a big difference for text classification. In: SIGIR, Seattle, Washingthon. 1995:130–6.